Forensic Science International: Genetics 10 (2014) 23–32

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Progress toward an efficient panel of SNPs for ancestry inference

Kenneth K. Kidd [a,*], William C. Speed [a], Andrew J. Pakstis [a], Manohar R. Furtado [b,1], Rixun Fang [b], Abeer Madbouly [c], Martin Maiers [c], Mridu Middha [c,2], Françoise R. Friedlaender [d], Judith R. Kidd [a]

[a] Department of Genetics, Yale University School of Medicine, New Haven, CT 06520-8005, USA
[b] Life Technologies, 1149 Chess Drive, Foster City, CA 94404, USA
[c] Bioinformatics Research, National Marrow Donor Program, 3001 Broadway Street NE, Suite 100, Minneapolis, MN 55413, USA
[d] 23 Hunting Ridge Road, Sharon, CT 06069, USA

ARTICLE INFO

Article history:
Received 8 August 2013
Received in revised form 3 January 2014
Accepted 7 January 2014

Keywords:
SNP
Ancestry
Population differentiation
AIM
Structure

ABSTRACT

Many panels of ancestry informative single nucleotide polymorphisms have been proposed in recent years for various purposes including detecting stratification in biomedical studies and determining an individual's ancestry in a forensic context. All of the panels have limitations in their generality and efficiency for routine forensic work. Some panels have used only a few populations to validate them. Some panels are based on very large numbers of SNPs thereby limiting the ability of others to test different populations. We have been working toward an efficient and globally useful panel of ancestry informative markers that is comprised of a small number of highly informative SNPs. We have developed a panel of 55 SNPs analyzed on 73 populations from around the world. We present the details of the panel and discuss its strengths and limitations.

1. Introduction

The many published sets of ancestry informative markers (AIMs) over the last decade [1–18] and papers on methods to identify AIMs [6,19] attested to the importance of AIMs. These studies have mostly used SNPs or di-allelic insertion–deletion markers (InDels or DIPs) because the forensic STR markers are not especially powerful for ancestry inference [20,21]. SNP sets have been developed for various reasons: estimating admixture in individuals from populations known to be admixed, usually involving specific ancestral populations; distinguishing an individual's ancestral origins assuming no significant admixture involving distant populations; controlling for heterogeneous ancestry in clinical association studies. Forensic identification of ethnicity has been yet another reason for developing such sets of markers. The variety of population resources used to identify the ancestry informative SNPs has ranged from a few widely separated population samples in the HapMap to the HGDP-CEPH panel of 52 small population samples.

Very large numbers of markers will nearly always provide accurate discrimination for at least 6 or 7 geographic regions. However, most useful for forensics would be a small but efficient and robust set of markers that would provide excellent information on ancestry. We have previously identified a panel of SNPs that have both high heterozygosity globally and very low allele frequency variation around the world [22,23]. This panel is of great forensic value for individual identification but gives no information on ancestry. In contrast, an optimized panel of ancestry informative SNPs (AISNPs, a subset of AIMs in general) will need SNPs with large allele frequency differences among a very broad set of populations. A limitation of AIMs in general is that they cannot distinguish among populations not previously studied. Thus, individual ancestry estimation is problematic if a relevant ancestral population has not been included in the defining studies.

Our interest in AISNPs is forensics: we wish to identify a small number of SNPs that will be good for identifying the geographic/ethnic origin of an unknown sample. The origin estimated must have a high enough probability of being correct that the SNPs will provide a useful investigative tool. In a forensic context a small number of SNPs can mean lower costs and possibly faster

* Corresponding author at: Department of Genetics, Yale University School of Medicine, 333 Cedar Street, PO Box 208005, New Haven, CT 06520-8005, USA. Tel.: +1 203 785 2654; fax: +1 203 785 6568.
E-mail address: kenneth.kidd@yale.edu (K.K. Kidd).
[1] Current address: Biology for Global Good, 420 Ventura Place, San Ramon, CA, USA.
[2] Current address: Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN 55905, USA.

1872-4973
http://dx.doi.org/10.1016/j.fsigen.2014.01.002

turnaround. A small number of highly selected SNPs can be sufficient for accurate estimation of ancestry [24]. The search for optimal SNPs must use population samples that are representative of diverse geographical regions and have large enough sample sizes so that sampling errors are minimized. One must then identify those polymorphisms most able to distinguish among those populations. We have used enough different population samples that we have several samples from each major geographic region we are investigating and individual population sample sizes averaging 50 individuals. We have selected candidate SNPs using a wide variety of methods and sources. In this report we present our current set of 55 AISNPs that constitute an efficient panel for a global distinction of seven to eight biogeographic regions.

2. Methods

2.1. Strategy

We used many sources of data to identify potential AISNPs. We initially used the Applied Biosystems database of allele frequencies of four populations (Japanese, Chinese, Europeans, African Americans) for the TaqMan probes they sell. SNPs with a frequency range near 1.0 became candidates. Next we used the ~650,000 SNPs tested on the HGDP-CEPH panel of over 1000 individuals from 51 populations [25], as have others [12–17]. We also used data we collected for the same SNPs tested on 1300 additional individuals not present in the HGDP. These additional individuals increased the sample sizes for the populations we contributed to the HGDP and added additional populations. We used our own laboratory database of about 4000 polymorphic markers typed on from 44 to 56 populations consisting of a total of nearly 3000 individuals. Our laboratory database resulted from many different studies of allele frequency variation done for a variety of reasons, e.g., pharmacogenetics [26]. As they became available we screened other large datasets for promising candidate AISNPs.

We explored several approaches to selecting candidate SNPs, comparing them, and balancing the information a selection provided. Ultimately, the combination of approaches would have to be considered empiric. Many candidate SNPs initially had data on a small number of populations; we selected those sites that had the largest absolute frequency differences or the largest F_{st} values for further evaluation. They were tested on our initially available set of 44 populations. Combined analyses of published datasets is often impossible because different studies used different markers on different populations [26]. Two published panels are based on the HGDP data [27]: the set of 128 SNPs identified by Seldin's group [12] and the set of 41 SNPs identified by Nievergelt et al. [17]. These have no SNPs in common but can be analyzed together since the individuals studied are the same. To help overcome the general dearth of SNPs studied in common we analyzed the 128 SNPs from Seldin's group on our populations [28] and included data on our populations in the Nievergelt study. In both cases some SNPs had already been identified by us as good candidates; both studies also included other SNPs we had not previously identified as excellent candidates. All of the markers from those two studies were included in the set of several hundred candidate AISNPs that were typed on the remaining samples in our lab to complete a comprehensive dataset with no missing population-SNP data points. The global coverage of our several hundred candidate AISNPs consisted of 63 populations with a total of 3071 individuals (see list in Supplemental Table S1).

2.2. Balancing information

It is important to balance the selection of SNPs such that the information from different SNPs assures that different geographical regions of the world are robustly distinguishable [14,29]. For example, a random selection of SNPs with high global F_{st} will have a large excess of SNPs with allele frequencies distinguishing African populations from populations in the rest of the world, a dichotomy that can outweigh most other distinctions among populations. We used several methods to balance the SNP selection. Our approach to identifying highly informative AIMs is analogous to other approaches [14,29] but differed from them in that we used all $(63 \times 62)/2$ pairwise comparisons of our 63 populations to identify SNPs with the largest pairwise allele frequency differences. This allowed us to identify markers especially useful for discriminating among populations from many different biogeographic regions. In contrast, other studies often focused on comparing more restricted predefined regions appropriate for each specific research question. Heatmaps of the candidate gene allele frequencies helped by

Table 1
The 55 AISNPs.

dbSNP rs#	Chr	Build 37 nt position	73-population F_{st}
rs3737576	1	101,709,563	0.44
rs7554936	1	151,122,489	0.39
rs2814778	1	159,174,683	0.82
rs798443	2	7,968,275	0.34
rs1876482	2	17,362,568	0.75
rs1834619	2	17,901,485	0.50
rs3827760	2	109,513,601	0.71
rs260690	2	109,579,738	0.49
rs6754311	2	136,707,982	0.41
rs10497191	2	158,667,217	0.54
rs12498138	3	121,459,589	0.48
rs4833103	4	38,815,502	0.37
rs1229984	4	100,239,319	0.43
rs3811801	4	100,244,319	0.45
rs7657799	4	105,375,423	0.44
rs16891982	5	33,951,693	0.69
rs7722456	5	170,202,984	0.20
rs870347	6	6,845,035	0.35
rs3823159	6	136,482,727	0.50
rs192655	7	90,518,278	0.21
rs917115	8	28,172,586	0.35
rs1462906	8	31,896,592	0.54
rs6990312	8	110,602,317	0.34
rs2196051	8	122,124,302	0.43
rs1871534	8	145,639,681	0.48
rs3814134	9	127,267,689	0.47
rs4918664	10	94,921,065	0.53
rs174570	11	61,597,212	0.51
rs1079597	11	113,296,286	0.16
rs2238151	12	112,211,833	0.36
rs671	12	112,241,766	0.22
rs7997709	13	34,847,737	0.37
rs1572018	13	41,715,282	0.41
rs2166624	13	42,579,985	0.30
rs7326934	13	49,070,512	0.54
rs9522149	13	111,827,167	0.44
rs200354	14	99,375,321	0.32
rs1800414	15	28,197,037	0.57
rs12913832	15	28,365,618	0.52
rs12439433	15	36,220,035	0.39
rs735480	15	45,152,371	0.39
rs1426654	15	48,426,484	0.73
rs459920	16	89,730,827	0.24
rs4411548	17	40,658,533	0.14
rs2593595	17	41,056,245	0.47
rs17642714	17	48,726,132	0.18
rs4471745	17	53,568,884	0.27
rs11652805	17	62,987,151	0.39
rs2042762	18	35,277,622	0.43
rs7226659	18	40,488,279	0.40
rs3916235	18	67,578,931	0.63
rs4891825	18	67,867,663	0.53
rs7251928	19	4,077,096	0.47
rs310644	20	62,159,504	0.58
rs2024566	22	41,697,338	0.31

graphically portraying redundancy in SNP information. Pairwise F_{st} calculations for each SNP across populations from different regions helped identify those SNPs best at certain distinctions, such as Europe vs. East Asia, so that the SNPs best at pairwise distinctions were used in the balancing. We also employed STRUCTURE [30] as one first-pass method of identifying the SNPs that differentiated most between the clusters identified. After a considerable amount of testing alternative sets of SNPs and switching individual SNPs in and out, we present a more efficient provisional panel of 55 AIMs. Once we had identified our set of 55 AISNPs on our 63 populations, we extracted the data for 813 individuals from the 1000 Genomes populations. The resulting data include 73 populations and 3884 individuals.

2.3. Laboratory

The 63 population samples from our laboratory were typed for all SNPs by TaqMan SNP Genotyping Assays® (Applied Biosystems, Foster City, California, USA) in three microliter reactions following the manufacturer's instructions. The genotypes of the samples in the 1000 Genomes Project were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/. Overall, missing genotypes account for 1.5% of the total, with no SNP exceeding 4% missing genotypes in the 3884 individuals.

2.4. Statistics

F_{st} was calculated for the allele frequencies using the formula of Wright with no modification for sample size variation among the population samples [28]. We did not also use Rosenberg's I_n statistic because it was shown to be highly correlated with F_{st} [24]. We used both the overall F_{st} in selecting candidate SNPs and the pairwise F_{st} in balancing the panel to include SNPs informative for different distinctions among populations. Heatmaps were calculated using the public program in R. Principal components analysis (PCA) of population sample allele frequencies used XLSTAT (version 2009.4.07; Addinsoft SARL, http://www.xlstat.com/en/company/)). MDS, using XLSTAT on the dataset of 63 populations, was used to illustrate the diversity of SNP information.

STRUCTURE (version 2.3.4; software freely available at http://pritch.bsd.uchicago.edu/structure.html) [30–32] was also used to evaluate and visualize the degree to which sets of sites distinguish among the populations. The various analyses used a burn-in of 20,000 followed by 10,000 iterations with a model of correlated allele frequencies specified. Specific solutions were plotted using DISTRUCT 1.1 (free software downloaded from http://rosenberglab.bioinformatics.med.umich.edu/distruct.html) [33]. For the final set of 55 AISNPS ten replicates at each of the "K" levels 2–6 and 20 replicates at $K = 7–8$ were evaluated using CLUMPP (free software downloaded from http://rosenberglab.bioinformatics.med.umich.edu/clumpp.html) [34]. The matrix of pairwise similarities among replicate runs was employed to identify different overall patterns based on high G values among runs with the "same" pattern and lower values for runs with different patterns.

Calculation of likelihoods of ancestry for selected individuals used the function in FROG-kb http://frog.med.yale.edu for the Kidd Lab 55 AISNP panel described in this paper. For each population the calculation is simply the product of the frequencies of the genotypes of the input individual across all 55 loci. In the output the populations are ranked from highest to lowest likelihood.

3. Results and discussion

The final list of 55 AISNPs is given in Table 1. The allele frequencies are available in ALFRED for these 73 populations and any other populations that have data available in ALFRED. The data can be retrieved under the individual rs-numbers or through the "SNP Sets" menu as "KiddLab Set of 55 AISNPs". There were no significant deviations beyond chance levels for Hardy-Weinberg ratios given the $55 \times 73 = 4015$ tests. Fig. 1 compares the

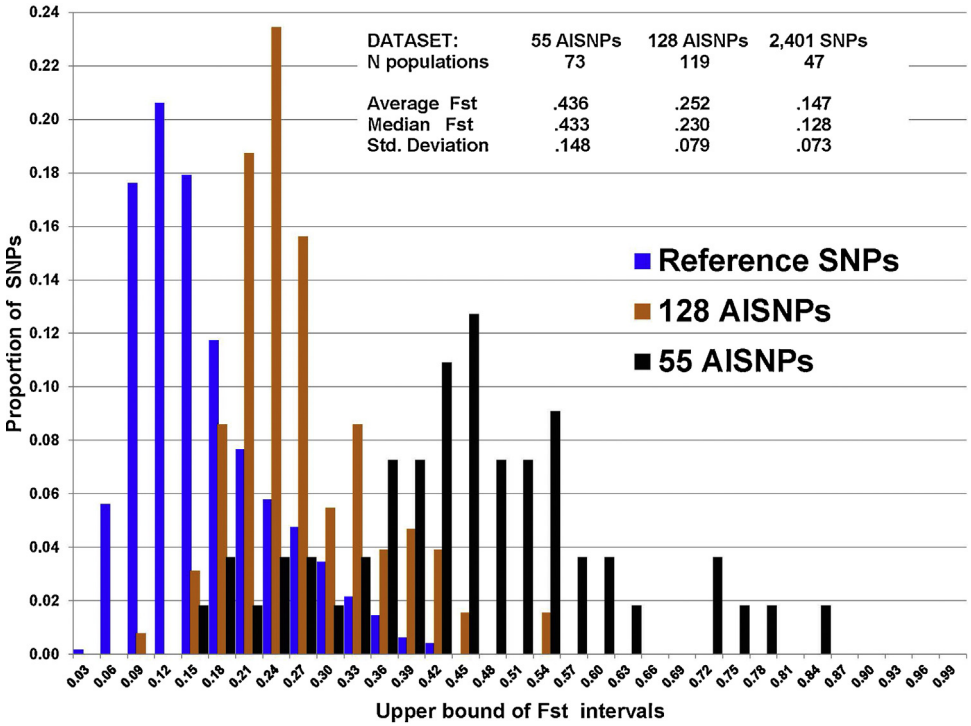

Fig. 1. Comparison of F_{st} distributions. Two previously published distributions (Kidd et al., [28]) are compared to the distribution for the set of 55 AISNPs. The two previous distributions are based on a reference set of SNPs typed on the Kidd Lab populations and on the Seldin group's set of 128 Ancestry Informative SNPs typed on a larger set of populations including the Kidd Lab populations. Because all three sets include the basic 47 Kidd Lab populations, the additional and different populations in the two larger studies are not sufficient to invalidate the marked differences in the distributions.

distributions of F_{st} for these 55 AISNPs with those for two other sets of markers: an essentially random set of SNPs [22] and the published set of 128 AISNPs [12,13,28]. Though the three distributions are based on different numbers of populations, many population samples occur in all three data sets and the geographic ranges of populations are the same. On average, we are dealing with a set of SNPs with greater global variation than the 128 AISNPs. The Nievergelt et al. [17] AISNPs, based on available population data in ALFRED, have a mean and median F_{st} of 0.36, intermediate between the 128- and 55-AISNP panels.

The heatmap in Fig. 2 is based on the population allele frequencies for the 55 AISNPs. It allows a very quick visualization of (1) the relationship of each SNP in the data set to the others, and (2) of how each SNP contributes to distinguishing among populations. The heatmap shows the relationships of the SNPs and of the populations graphically in the marginal dendrograms. The heatmap also allows a determination of how these individual markers contribute to the differentiation of the specific populations analyzed. The several higher branchings of the SNP dendogram indicate that diverse patterns of allele frequency variation occur among these 55 AIMs.

STRUCTURE is useful for displaying how individual genotypes for a set of AIMs segregate individuals into approximately Mendelian populations. In the most likely STRUCTURE run at $K = 8$ the 3884 individuals in this study are assigned to seven distinct clusters in which most individuals in most populations fall into a single clusters (Fig. 3). At $K = 8$ the results for most individuals in most populations are essentially unaltered from the pattern at $K = 7$ (not shown) but a complex "admixture" pattern is introduced for the European populations. PCA on the allele frequencies in the populations shows four distinct groupings of populations based on the first 3 components (Fig. 4): a highly

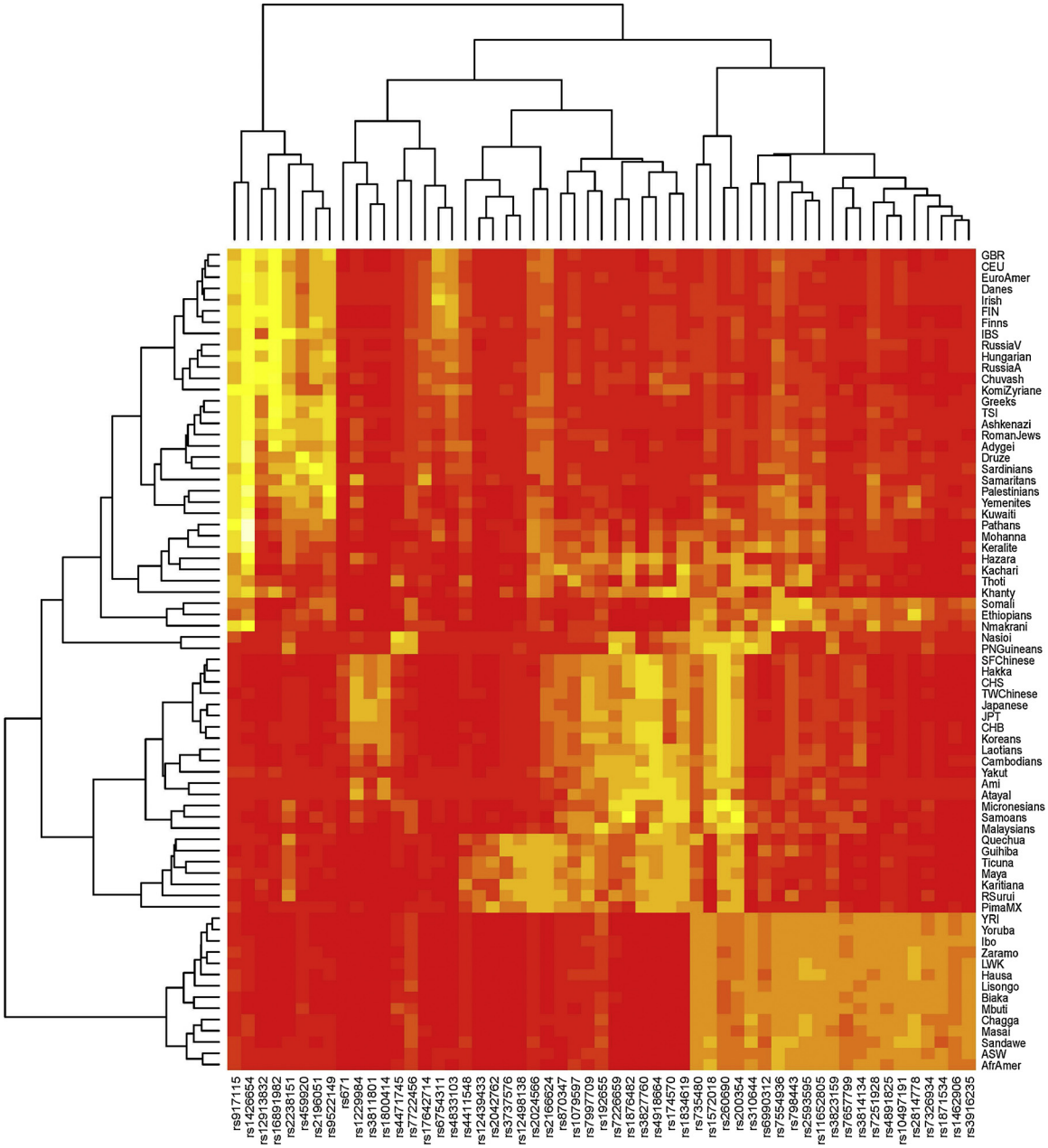

Fig. 2. The heatmap of the clustering of the 73 populations and the 55 AISNPs. The upper left block represents Europe through South Central Asia. The large middle block represents East Asia and below that the Native Americans. The bottom right block represents Africa. Clearly, different SNPs contribute differently to population distinctions and one view of the relationships is given by the lengths of the branches in the dendograms.

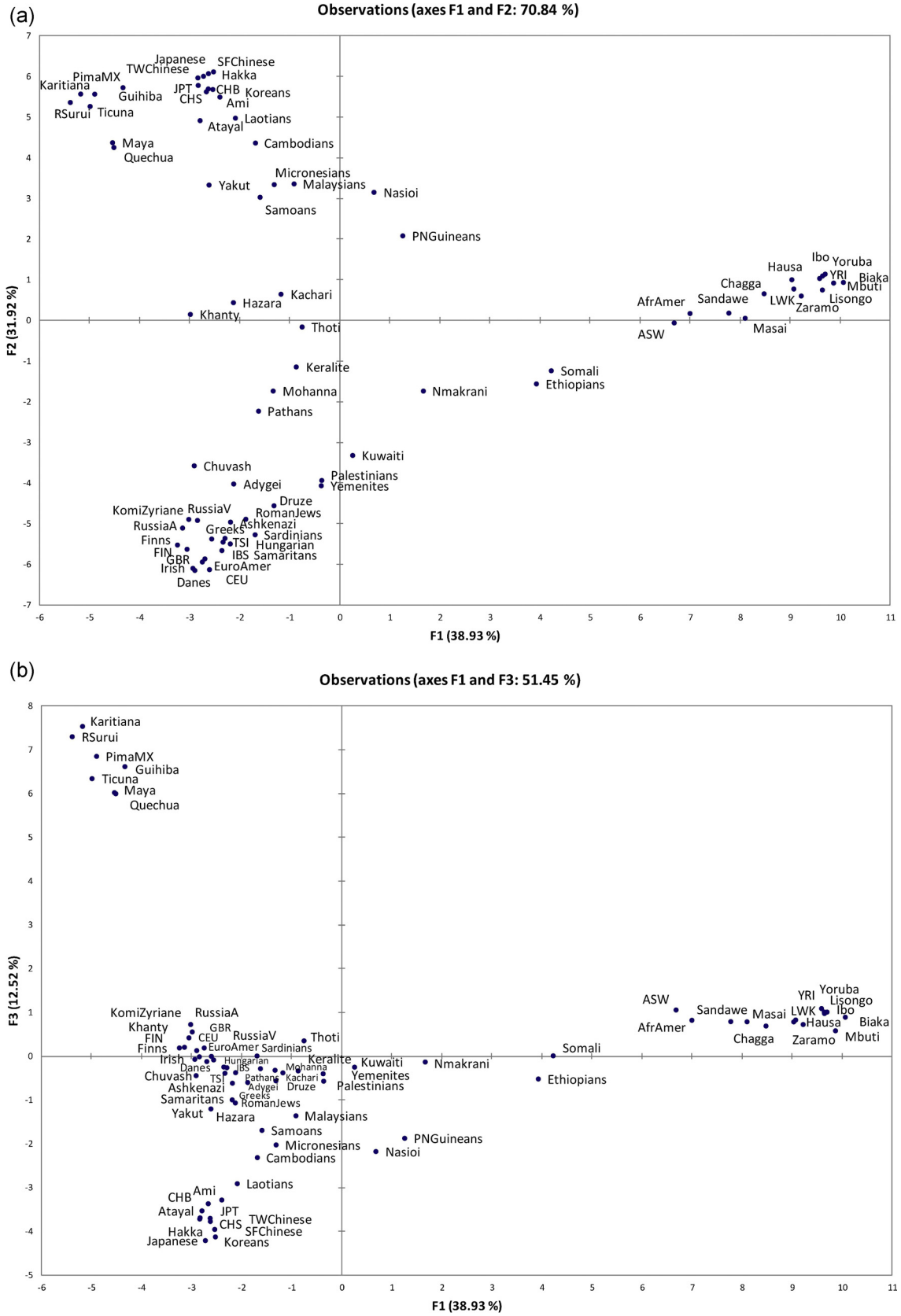

Fig. 3. Principal Component Analysis of the 73 populations using the 55 AISNPs. (a) The first PC accounts for 38.9% of the variance and primarily separates African populations from the rest of the world. The second PC accounts for 31.9% of the variance and primarily separates Europe from East Asia and the Americas. The two components account for 70.8% of the variance. (b) The third PC accounts for 12.5% of the variance and completely separates the American Indians from the East Asians.

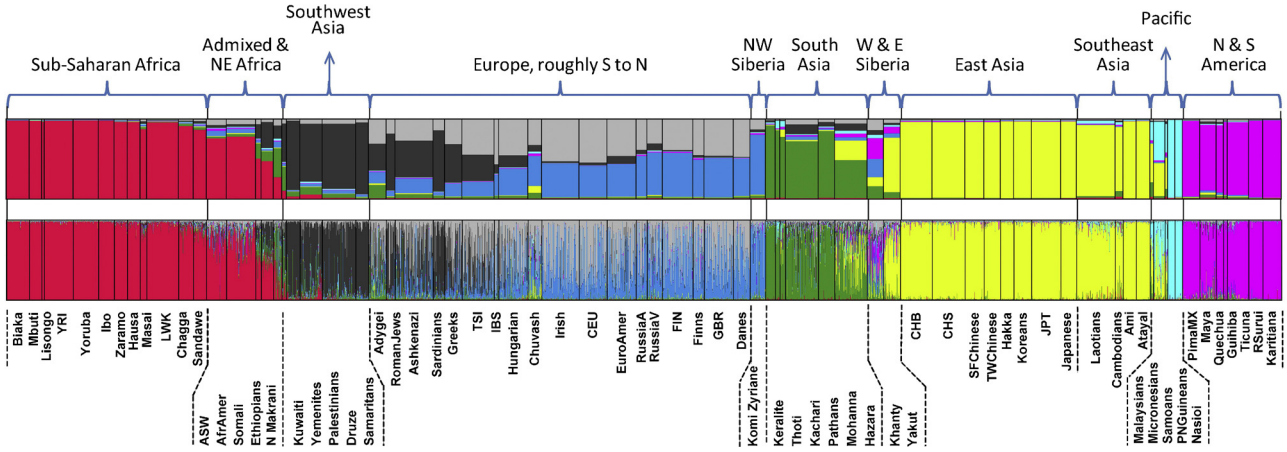

Fig. 4. The most likely of the 20 STRUCTURE analyses at $K = 8$ for the full dataset. The results are plotted as the average assignments for each population and as the individual assignments. A cline is evident for the Mediterranean populations between the populations in Southwest Asia and those in central and northern Europe. We note also that all of the European populations have been estimated to be admixed between two clusters (illustrated in gray and blue) not otherwise present. This likely relates to the inherent Mendelian segregation for most of the "European specific" markers.

distributed African group, a more tightly clustered East Asian group, a modestly clustered Native American group, and a European-Southwest Asian group. This pattern reflects the geographic clustering of the majority of the populations being studied: the geographically intermediate populations tend to be placed in more intermediate positions. The African populations show a West to East cline toward the non-African populations. Taken together, the heatmap and the STRUCTURE analyses show that clusters exist in which several populations are essentially indistinguishable. These analyses demonstrate that information exists on ancestral origins of individuals, but does not obviously indicate how strongly the clusters differ in a statistical sense.

Although STRUCTURE allows evaluation of potential AISNPs, it is cumbersome to use and not particularly useful in our effort to identify as small a set of SNPs as possible while still defining multiple geographic regions of origin. The empiric approach using multiple methods as described above produced surprisingly good results. The value of an ancestry panel depends on how accurately a likelihood function determines ancestry of an individual. That accuracy will depend on the specific ancestry of the individual, the reference populations available for comparison, and the particular set of SNPs. We illustrate this by estimating the population assignments of six individuals not otherwise in the study: two Hungarians, two Druze, and two Mongolians. The Hungarian and Druze individuals were not included in the reference data or used to select the panel of SNPs but are related to individuals in those datasets. The two unrelated Mongolian individuals are recruits from among the students of the Health Sciences University of Mongolia in Ulaan-Baatar; no reference population data for Mongolia are available for calculations. For all six individuals we have used the functions in FROG-kb [35] to calculate the likelihoods of the individual originating from each of our 63 populations. In Tables 2–4 we list the likelihoods and likelihood ratios for the top 20 populations for each of the six individuals. The likelihoods are graphed in supplemental figures S3 through S5 in numeric order for all 63 populations already incorporated in FROG-kb.

These results illustrate several points. In their analysis of Spanish vs. Moroccan ancestry, Phillips et al. [36] showed that likelihood of ancestral assignment to the two populations differed among individuals and that a few individuals were misclassified or not classified with statistical significance. With 63 reference

Table 2
FROGkb output for 2 Hungarian individuals – 55 AISNP panel.

Hungarian A			Hungarian B		
Population	Probability of genotype	Likelihood ratio	Population	Probability of genotype	Likelihood ratio
Samaritans	1.7E−14		Hungarians	3.0E−14	
Roman Jews	1.3E−14	1.3E+00	Russians Vol.	1.1E−14	2.7E+00
Ashkenazi	6.3E−15	2.7E+00	Finns	7.9E−15	3.8E+00
Druze	4.1E−16	4.1E+01	EuroMixed	5.3E−15	5.6E+00
Hungarians	3.0E−16	5.6E+01	Chuvash	4.7E−15	6.4E+00
Russians Arch.	2.6E−16	6.5E+01	Komi Zyrian	3.6E−15	8.3E+00
Greeks	1.2E−16	1.4E+02	Russians Arch.	2.8E−15	1.1E+01
Russians Vol.	7.7E−17	2.2E+02	Danes	1.5E−15	2.1E+01
Finns	1.6E−17	1.0E+03	Ashkenazi	1.1E−15	2.7E+01
Toscani	9.9E−18	1.7E+03	Irish	6.1E−16	4.9E+01
Chuvash	4.4E−18	3.8E+03	Adygei	3.2E−16	9.5E+01
Danes	3.6E−18	4.6E+03	Greeks	3.0E−16	9.9E+01
EuroMixed	3.6E−18	4.7E+03	Roman Jews	9.2E−17	3.3E+02
Arabs Palestine	2.9E−18	5.8E+03	Sardinians	1.6E−17	1.8E+03
Sardinians	1.4E−18	1.2E+04	Toscani	1.3E−17	2.3E+03
Adygei	1.1E−18	1.6E+04	Pathans	1.9E−18	1.6E+04
Irish	6.3E−19	2.7E+04	Arabs Palestine	6.0E−19	5.0E+04
Komi Zyrian	3.3E−19	5.0E+04	Kuwaiti	2.1E−19	1.4E+05
Yemenite Jews	1.4E−19	1.2E+05	Druze	8.4E−21	3.6E+06
Pathans	4.7E−21	3.6E+06	N Makrani	5.5E−21	5.4E+06

Table 3
FROGkb output for 2 Druze individuals – 55 AISNP panel.

Druze A			Druze B		
Population	Probability of genotype	Likelihood ratio	Population	Probability of genotype	Likelihood ratio
Toscani	5.8E−15		Adygei	8.5E−12	
Greeks	4.5E−15	1.3E+00	Toscani	2.2E−12	3.9E+00
Adygei	3.6E−15	1.6E+00	Arabs Palestine	2.0E−12	4.2E+00
Druze	2.5E−15	2.3E+00	Sardinians	2.0E−12	4.2E+00
Ashkenazi	9.1E−16	6.4E+00	Ashkenazi	4.4E−13	1.9E+01
EuroMixed	4.5E−16	1.3E+01	Greeks	3.3E−13	2.6E+01
Roman Jews	4.3E−16	1.4E+01	Druze	2.7E−13	3.1E+01
Arabs Palestine	2.3E−16	2.5E+01	Roman Jews	2.1E−13	4.0E+01
Pathans	1.6E−16	3.7E+01	N Makrani	9.1E−14	9.3E+01
Russians Vol.	1.2E−16	4.7E+01	EuroMixed	8.5E−14	1.0E+02
Hungarians	1.1E−16	5.2E+01	Pathans	6.0E−14	1.4E+02
Chuvash	7.7E−17	7.5E+01	Hungarians	4.9E−14	1.7E+02
Yemenite Jews	5.1E−17	1.1E+02	Yemenite Jews	4.4E−14	1.9E+02
Sardinians	1.4E−17	4.1E+02	Kuwaiti	4.0E−14	2.1E+02
Kuwaiti	8.1E−18	7.1E+02	Mohanna	3.4E−14	2.5E+02
Danes	6.0E−18	9.6E+02	Russians Vol.	1.1E−14	7.6E+02
Mohanna	3.1E−18	1.9E+03	Keralites	7.6E−15	1.1E+03
Komi Zyrian	1.8E−18	3.3E+03	Danes	6.0E−15	1.4E+03
N Makrani	8.5E−19	6.8E+03	Chuvash	3.6E−15	2.4E+03
Finns	3.9E−19	1.5E+04	Irish	1.2E−15	6.8E+03

populations many more options for "miss-assignment" are possible. Because of Mendelian segregation some individuals in a population may have genotypes that are more likely to occur in a population other than the population of origin. However, the other populations that have higher or similar likelihoods of origin are generally from the same or a nearby region. For the two Druze individuals the other high-ranking populations are generally Mediterranean. The two Hungarians show much different sets of high-ranking populations of origin and the results could be interpreted as Hungarian A having significant Jewish ancestry, an entirely plausible result given known European history. Finally, the two Mongolian individuals have neither a "correct" ancestral population nor any geographically close populations among the reference populations available for assignment. They show quite different rankings of Asian populations and illustrate the high inherent uncertainty in estimating the ancestry of an individual originating from a poorly represented region of the world. Thus, using a likelihood function such as implemented for this panel in FROG-kb [35] cannot be expected to identify routinely the specific population from which an individual originates. Rather, the best resolution one can be reasonably confident of is that the cluster of populations (as seen in Fig. 4) an individual belongs to will be identified but not necessarily with high statistical significance.

To distinguish among populations from many different regions of the world requires SNPs that have a variety of patterns of allele frequencies around the world. We have used MDS of the SNPs to evaluate the diversity of the 55 SNPs (Fig. 5). The variety of patterns of allele frequency variation is reflected in the SNPs' dispersion on the MDS plot. The only very tight cluster occurs at the bottom of the figure and represents several SNPs that provide a primarily Africa vs. the rest of the world picture. Several SNPs are highlighted in Fig. 5. Their frequency patterns are illustrated in other figures. Fig. 6 shows four SNPs with relatively simple patterns; each differentiates a single geographic region. In combination, however, the set of four clearly distinguishes the Pacific populations and the East African populations. Figures S1 and S2 in supplementary material illustrate the allele frequency patterns of the other SNPs highlighted in Fig. 5.

Table 4
FROGkb output for 2 Mongolian individuals – 55 AISNP panel.

Mongolian A			Mongolian B		
Population	Probability of genotype	Likelihood ratio	Population	Probability of genotype	Likelihood ratio
Yakut	8.1E−14		Chinese TW	2.0E−14	
Cambodians	6.2E−15	1.3E+01	Laotians	1.0E−14	2.0E+00
Hazara	2.8E−15	2.9E+01	Japanese	1.9E−15	1.0E+01
Kachari	1.2E−15	6.6E+01	Chinese SF	1.9E−15	1.0E+01
Laotians	8.8E−17	9.2E+02	Ami	1.7E−15	1.1E+01
Malaysians	6.0E−17	1.4E+03	Koreans	3.4E−16	5.9E+01
Chinese TW	3.3E−17	2.4E+03	Hakka	2.9E−16	6.9E+01
Chinese SF	1.3E−17	6.3E+03	Cambodians	1.8E−16	1.1E+02
Koreans	3.4E−18	2.4E+04	Samoans	9.3E−17	2.1E+02
Khanty	2.4E−18	3.4E+04	Atayal	5.1E−17	3.9E+02
Micronesians	2.4E−18	3.4E+04	Yakut	2.3E−17	8.8E+02
Hakka	4.9E−19	1.7E+05	Micronesians	1.2E−17	1.6E+03
Maya	2.9E−19	2.8E+05	Hazara	6.7E−18	3.0E+03
Samoans	9.3E−20	8.8E+05	Malaysians	2.1E−19	9.4E+04
Japanese	4.1E−20	2.0E+06	Kachari	4.3E−24	4.7E+09
Quechua	3.5E−20	2.3E+06	Khanty	6.0E−25	3.3E+10
Ami	2.5E−20	3.3E+06	Thoti	7.8E−26	2.6E+11
Pathans	2.7E−21	3.0E+07	Quechua	1.3E−26	1.5E+12
Ticuna	2.6E−21	3.2E+07	Pathans	2.9E−27	6.9E+12
Thoti	1.5E−21	5.5E+07	Chuvash	1.3E−27	1.5E+13

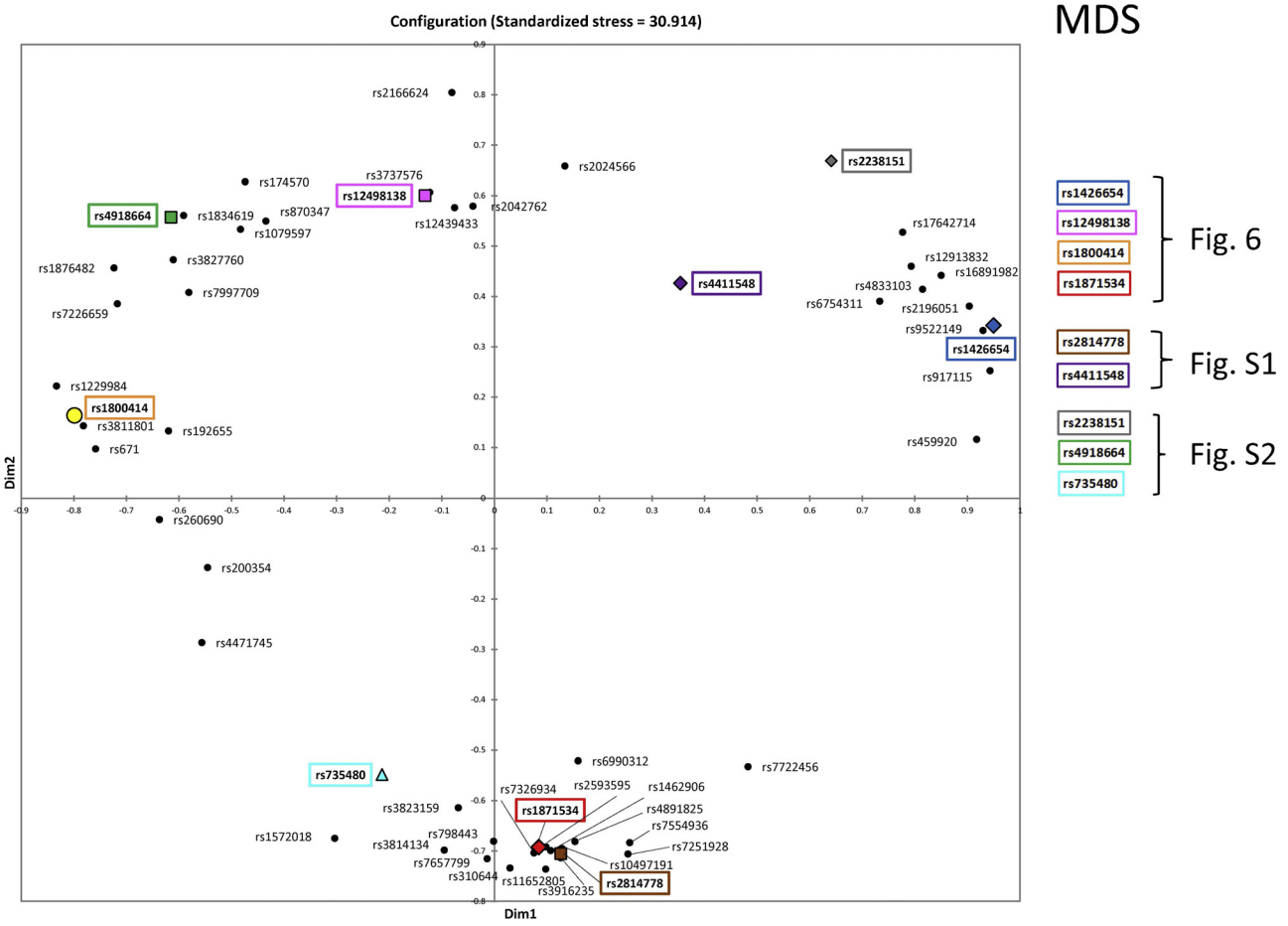

Fig. 5. An MDS plot of SNPs based on the pairwise correlations of SNP allele frequencies across all 73 populations. The dispersion of the SNPs in the plot reflects the variety of patterns of allele frequency variation shown by the different SNPs. Some specific examples are shown in Fig. 6 and in Supplementary Figures S1 and S2.

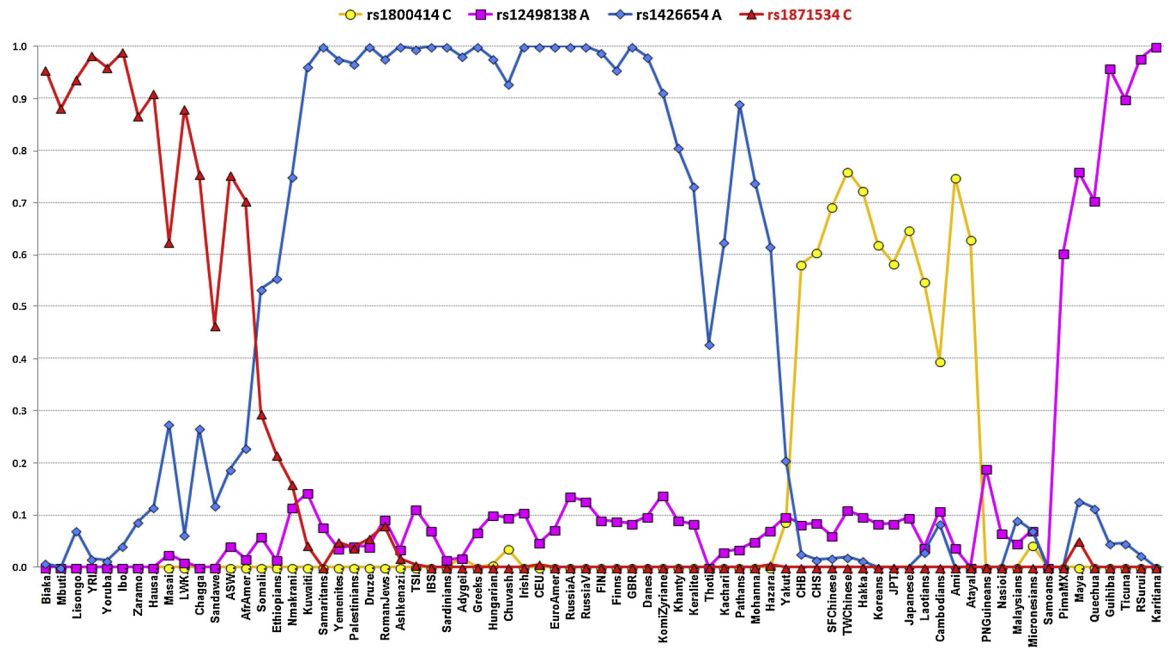

Fig. 6. The allele frequency distributions for four SNPs that are highly differentiated in a single region of the world. These four SNPs are among those at the compass points in the MDS plot (Fig. 5). In addition to SNPs highly differentiating each of four biogeographic regions: rs2814778 in the South distinguishing Africa; rs1426654 in the East distinguishing Europe; rs12498138 in the North distinguishing Native Americans; and rs1800414 in the West distinguishing East Asians. Three of these SNPs have ranges essentially spanning zero to one. Geographically intermediate populations often have intermediate frequencies; we note especially the populations in Southwest Asia.

Can this panel of AISNPs be improved? Absolutely. Resolution of ancestry, especially for individuals from populations not represented in these 73, will likely be improved if more populations are typed for these SNPs. However, the greatest improvement will come from using "better" SNPs. The problem is finding SNPs that provide a clearer differentiation of certain populations or groups of populations without detracting from differentiation among some other populations. As noted in Kersbergen et al. [14], some SNPs simply add noise. We note that several of the SNPs that help differentiate European individuals from the rest of the world are not fixed for the Europe-specific allele. With genotype differences among individuals some individuals will tend to have the non-European alleles at more of the loci than other individuals. At higher K values the STRUCTURE analyses apparently use this Mendelian segregation to classify individuals in all European populations "randomly" into two or three different clusters, as seen in Fig. 4. In general, even if a SNP has extreme frequency variation between, say, East Asians and Native Americans, but the frequencies in Europe and Southwest and South Asia are all intermediate with no population distinguishing pattern, that SNP is adding noise to the differentiation of those populations. The SNP with the lowest F_{st} in these 73 populations, rs4411548, illustrates exactly that situation (Supplemental Figure S1). The frequency of one allele is near zero in East Asian and Pacific populations and ranges from 19% to 86% in Native Americans. In contrast, that allele ranges from 2% to 45%, with most other populations between 12% and 30%, in Africans, Europeans, and Southwest and South Central Asians. We have found that it is difficult to find additional SNPs that differentiate populations both globally and within regions while, at the same time, minimizing the total number of SNPs. An alternative approach that we are considering is a second tier of SNPs that are good within a region but not necessarily good, or as good as existing AISNPs, for global differentiation. We are currently working on one such second tier of AISNPs for the eastern half of Asia. Phillips et al. [18] have proposed such a regional panel focused on distinguishing European from South Asian populations. Another approach we are pursuing is the use of haplotypes comprised of molecularly close SNPs [37,38].

4. Conclusions

The variety of approaches we have used to optimize a set of ancestry informative SNPs all have value but none seems sufficient. The final test is how well the panel will rank the potential populations of ancestry in a likelihood context. While the current likelihood calculations in FROG-kb do not explicitly allow admixed ancestry involving different biogeographic regions, the possibility of admixed ancestry raises a caveat in use of any statistic with any panel of AIMs. Admixed ancestry cannot be estimated accurately unless the ancestral populations are represented among the reference populations.

While we note that improvements will likely be possible for this panel, our analyses show it is a very good first tier panel for identifying major geographic regions for the ancestry of an individual. Future tests of the robustness of this panel will require that additional populations be tested for these SNPs to determine how well the panel resolves ancestries for individuals from populations that are in poorly represented biogeographic regions and populations intermediate to the existing 73 population samples. Future improvement in resolution of ancestry among populations poorly differentiated by these 55 AISNPs will require searching for appropriate additional SNPs.

Conflict of interest

R. Fang works for Applied Biosystems/Life Technologies. The authors declare no other conflicts of interest.

Ethical approval

All samples were obtained with informed consent for studies of gene frequency variation under a protocol approved by the Yale IRB and by additional approved institutional protocols and government approvals as relevant in the various countries of origin.

Acknowledgments

This work was funded primarily by NIJ Grants 2010-DN-BX-K225 and 2010-DN-BX-K226 to KKK awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. Points of view in this presentation are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. We thank Eva Haigh for excellent technical help. We thank Drs. Jane Brissenden, Baigalmaa Evsanaa, Ariunaa Togtokh and Janet Roscoe for making the two Mongolian samples available. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies of gene frequency variation and to the many colleagues who helped us collect the samples. In addition, some of the cell lines were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, and the African American samples were obtained from the Coriell Institute for Medical Research, Camden, New Jersey.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2014.01.002.

References

[1] M.D. Shriver, E.J. Parra, S. Dios, C. Bonilla, H. Norton, C. Jovel, C. Pfaff, C. Jones, A. Massac, N. Cameron, et al., Skin pigmentation, biogeographical ancestry and admixture mapping, Hum. Genet. 112 (2003) 387–399.
[2] M.D. Shriver, R. Mei, E.J. Parra, V. Sonpar, I. Halder, S.A. Tishkoff, T.G. Schurr, S.I. Zhadanov, L.P. Osipova, T.D. Brutsaert, et al., Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation, Hum. Genomics 2 (2005) 81–89.
[3] H.E. Collins-Schramm, B. Chima, D.J. Operario, L.A. Criswell, M.F. Seldin, Markers informative for ancestry demonstrate consistent megabase-length linkage disequilibrium in the African American population, Hum. Genet. 113 (2003) 211–219.
[4] M. Bauchet, B. McEvoy, L.N. Pearson, E.E. Quillen, T. Sarkisian, K. Hovhannesyan, R. Deka, D.G. Bradley, M.D. Shriver, Measuring European population stratification with microarray genotype data, Am. J. Hum. Genet. 80 (2007) 948–956.
[5] O. Lao, K. van Duijn, P. Kersbergen, P. de Knijff, M. Kayser, Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry, Am. J. Hum. Genet. 78 (2006) 680–690.
[6] P. Paschou, E. Ziv, E.G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M.W. Mahoney, P. Drineas, PCA-correlated SNPs for structure identification in worldwide human populations, PLoS Genet. 3 (2007) 1672–1686.
[7] M.A. Enoch, P.H. Shen, K. Xu, C. Hodgkinson, D. Goldman, Using ancestry-informative markers to define populations and detect population stratification, J. Psychopharmacol. 20 (2006) 19–26.
[8] A.L. Price, N. Patterson, F. Yu, D.R. Cox, A. Waliszewska, G.J. McDonald, A. Tandon, C. Schirmer, J. Neubauer, G. Bedoya, et al., A genomewide admixture map for Latino populations, Am. J. Hum. Genet. 80 (2007) 1024–1036.
[9] X. Mao, A.W. Bigham, R. Mei, G. Gutierrez, K.M. Weiss, T.D. Brutsaert, F. Leon-Velarde, L.G. Moore, E. Vargas, P.M. McKeigue, et al., A genomewide admixture mapping panel for Hispanic/Latino populations, Am. J. Hum. Genet. 80 (2007) 1171–1178.
[10] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Alvarez-Dios, M. Calaza, M.C. de Cal, D. Ballard, M.V. Lareu, et al., Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (2007) 273–280.

[11] C. Tian, R. Kosoy, R. Nassir, A. Lee, P. Villoslada, L. Klareskog, L. Hammarström, H.J. Garchon, A.E. Pulver, M. Ransom, et al., European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups, Mol. Med. 15 (2009) 371–383.
[12] R. Kosoy, R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, et al., Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, Hum. Mutat. 30 (2009) 69–78.
[13] R. Nassir, R. Kosoy, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, et al., An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels, BMC Genet. 10 (2009) 39.
[14] P. Kersbergen, K. van Duijn, A.D. Kloosterman, J.T. den Dunnen, M. Kayser, P. de Knijff, Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans, BMC Genet. 10 (2009) 69.
[15] P. Qin, Z. Li, W. Jin, D. Lu, H. Lou, J. Shen, L. Jin, Y. Shi, S. Xu, A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese, Eur. J. Hum. Genet. 22 (2014) 248–253. , http://dx.doi.org/10.1038/ejhg.2013.111.
[16] O. Levran, O. Awolesi, P.H. Shen, M. Adelson, M.J. Kreek, Estimating ancestral proportions in a multi-ethnic US sample: implications for studies of admixed populations, Hum. Genomics 6 (2012) 2.
[17] C.M. Nievergelt, A.X. Maihofer, T. Shekhtman, O. Libiger, X. Wang, K.K. Kidd, J.R. Kidd, Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel, Invest. Genet. 4 (2013) 13.
[18] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, Á. Carracedo, P.M. Schneider, et al., Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359–366.
[19] S. Amirisetty, G.K. Hershey, T.M. Baye, Ancestry, SNPminer: a bioinformatics tool to retrieve and develop ancestry informative SNP panels, Genomics 100 (2012) 57–63.
[20] C. Phillips, L. Fernandez-Formoso, M. Gelabert-Besada, M. Garcia-Magariños, C. Santos, M. Fondevila, A. Carracedo, M.V. Lareu, Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing, Electrophoresis 34 (2013) 1151–1162.
[21] N.M. Silva, L. Pereira, E.S. Poloni, M. Currat, Human neutral genetic variation and forensic STR data, PLoS ONE 7 (2012) e49666.
[22] A.J. Pakstis, W.C. Speed, R. Fang, F.C.L. Hyland, M.R. Furtado, J.R. Kidd, K.K. Kidd, SNPS for a universal individual identification panel, Hum. Genet. 127 (2010) 315–324.
[23] K.K. Kidd, J.R. Kidd, W.C. Speed, R. Fang, M.R. Furtado, F.C. Hyland, A.J. Pakstis, Expanding data and resources for forensic use of SNPs in individual identification, Forensic Sci. Int. Genet. 6 (2012) 646–652.
[24] L. Ding, H. Wiener, T. Abebe, M. Altaye, R.C. Go, C. Kercsmar, G. Grabowski, L.J. Martin, G.K. Khurana Hershey, R. Chakorborty, et al., Comparison of measures of marker informativeness for ancestry and admixture mapping, BMC Genomics 12 (2011) 622.
[25] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, et al., Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100–11045.
[26] W.C. Speed, S.P. Kang, D.P. Tuck, L.N. Harris, K.K. Kidd, Global variation in CYP2C8-CYP2C9 functional haplotypes, Pharmacogenomics J. 9 (2009) 283–290.
[27] K.K. Kidd, W.C. Speed, A.J. Pakstis, J.R. Kidd, The search for better markers for forensic ancestry inference, in: Proceedings of the 22nd International Symposium on Human Identification; Sponsored by Promega Corp, 2011.
[28] J.R. Kidd, F.R. Friedlaender, W.C. Speed, A.J. Pakstis, F.M. De La Vega, K.K. Kidd, Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples, Invest. Genet. 2 (2011) 1.
[29] J.M. Galanter, J.C. Fernandez-Lopez, C.R. Gignoux, J. Barnholtz-Sloan, C. Fernandez-Rozadilla, M. Via, A. Hidalgo-Miranda, et al., Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas, PLoS Genet. 8 (3) (2012) e1002554.
[30] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[31] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet. 73 (2003) 1402–1422.
[32] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics 164 (2003) 1567–1587.
[33] N.A. Rosenberg, distruct: a program for the graphical display of population structure, Mol. Ecol. Notes 4 (1) (2004) 137–138.
[34] M. Jakobsson, N.A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, Bioinformatics 23 (2007) 1801–1806.
[35] H. Rajeevan, U. Soundararajan, A.J. Pakstis, K.K. Kidd, Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb, Invest. Genet. 3 (2012) 18.
[36] C. Phillips, L. Prieto, M. Fondevila, A. Salas, A. Gomez-Tato, J. Alvarez-Dios, A. Alonso, A. Blanco-Verea, M. Brion, M. Montesino, A. Carracedo, M.V. Lareu, Ancestry analysis in the 11-M Madrid bomb attack investigation, PLoS ONE 4 (8) (2009) e6583.
[37] A.J. Pakstis, R. Fang, M.R. Furtado, J.R. Kidd, K.K. Kidd, Mini-haplotypes as lineage informative SNPs (LISNPs) and ancestry inference SNPs (AISNPs), Eur. J. Hum. Genet. 20 (2012) 1148–1154.
[38] K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Lagace, J. Chang, S. Wootton, N. Ihuegbu, Microhaplotype loci are a powerful new type of forensic marker, Forensic Sci. Int.: Genet. Suppl. Ser. 4 (2013) e123–e124.

Electronic resources cited

ALFRED: http://alfred.med.yale.edu
FROG-kb: http://frog.med.yale.edu
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/
HapMap: http://hapmap.ncbi.nlm.nih.gov/
1000 Genomes project: http://www.1000genomes.org
AISNP set in ALFRED: http://alfred.med.yale.edu/alfred/selectedSnpSet.asp?setId=261